# Rapid response data-driven reconstructions for storm surge around New Zealand.

Tausia J[1], Delaux S[2], Camus P[1], Rueda A[1], Mendez F[1], Bryan K. R.[3], Perez J[2], Costa C. G. R.[2], Zyngfogel R[4], and Cofino A[5]

[1]Geomatic and Oceanographic Engineering Group, University of Cantabria, Spain
[2]Meteorological Service of New Zealand, New Zealand
[3]University of Waikato, Hamilton Waikato, New Zealand
[4]Calypso science, New Plymouth, New Zealand
[5]Consejo Superior de Investigaciones Científicas, Instituto de Física de Cantabria, Spain

**Correspondence:** Javier Tausia (tausiaj@unican.es)

**Abstract.** In conjunction with tides, storm surge is one major driver of coastal flooding associated with storm events. Because local inundation is strongly modulated by the local shape of the coastline and the bathymetric slope, accurate storm surge predictions using traditional numerical models require the use of very fine grids and are hence resource intensive. Therefore, the performance of a live prediction system based on such methods will likely be subject to a trade-off between prediction accuracy, prediction speed and cost.

This study explores the use of data driven methods as an alternative to numerical models to reconstruct the daily storm surge maximum levels along the entire coast of New Zealand. Firstly, several atmospheric predictors are utilized that incorporate different variables, time lags and spatial domains, using 3 statistical models, in a selected number of locations in New Zealand, to find the combination that optimizes the reconstruction. Finally, the storm surge daily maxima are reconstructed with the different statistical models along the entire coast, using the best performing predictor.

Results show very good performance for the best atmospheric predictor and statistical model, providing average values of 0.88 for the pearson correlation coefficient and 4.3cm for the root mean squared error metric (RMSE) (the average value for the RMSE in the 99% percentile is 8.2cm). For the Kling-Gupta Efficiency (KGE; incorporating 3 sub-metrics: correlation, bias term and variability term), which is the metric used to rank the models, the average value is 0.82.

Our results highlight the suitability of data driven models to simulate storm surge maximum levels, and prove the methodology is appropriate for finding a well performing atmospheric predictor that is able for reconstruct these values. Moreover, this methodology can be also applied to new variables, regions and problems, as there are no physical restrictions on the used predictors nor predictands.

**KEYWORDS -** Data-driven models, storm surge, atmospheric predictor, rapid reconstructions, New Zealand, coastline

## 1 Introduction

Flooding associated with storm surges is one of the most common natural hazards for coastal areas worldwide (Bell et al., 2000), and with over 15,000 km of coastline and around 150,000 people living in low-lying coastal areas, coastal inundation is also a major hazard to New Zealand (NZ). The cost to defend the associated buildings, infrastructure and assets is of the order of $10 billion (Ministry for the Environment (NZ), 2017, Preparing for Coastal Change, Publication number: ME 1335, 36pp.). With global sea level rise and the increase in the intensity and frequency of extreme weather events expected with climate change, the threat posed by coastal flooding is likely to become greater.

Storm surge is the rise of water level generated by wind and atmospheric pressure changes associated with tropical or extra-tropical (mid-latitude) storms, over and above the astronomical tide (AT), and the long-term signals such as the monthly mean sea level which contains the seasonal and inter-annual variability (Cid et al., 2017). Storm surge is one of the most critical components of coastal flooding and its magnitude has a large spatial variability (Bell and Goring, 1996).

The present work is focused on understanding storm surge behaviour in the two main islands of New Zealand: the North Island (or Te Ika-a-Maui), and the South Island (or Te Waipounamu), where AT accounts for 96% of the coastal energy, while the over-elevation associated with barometric pressure and wind effects, the effect of waves (Stephens et al., 2011), and the longer-term seasonal and inter-annual fluctuations account for the remaining 4% (Bell et al., 2000; Goring and Bell, 1996). Although storm surge around New Zealand, reaching just 0.8m maxima above mean sea level (Heath, 1979), is much lower than storm surge experienced in equatorial regions and high latitudes, it can still cause coastal flooding and exacerbate coastal erosion (Bell et al., 2000). For example, a flooding event that occurred in 1995 in the Thames Region, when peak storm surge overlapped with high AT, caused damage worth around 3–4 million dollars. In addition, during the spring and summer of 2017 and 2018, several large storms including ex-tropical cyclones Fehi, Gita, and Hola struck NZ, most of them coinciding with high perigean-spring tides, causing flooding to homes and damaging infrastructure. Other notable historical coastal flooding events in NZ occurred in January 2011, during cyclone Gisele in 1968 (de Lange and Gibb, 2000), May 1938 in the Hauraki Plains (Stephens et al., 2020) and during the great cyclone of 1936 (Brenstrum, 2000), but the spatial effects of these historical storms are not well recorded since not many sea-level gauges were in operation at those times (Stephens et al., 2019a).

Sea level forecasts usually use computationally expensive numerical models, which require running a model given the predicted atmospheric conditions every time a prediction is to be made, leading to substantial computational efforts (Wang et al., 2009; Siek, 2019; WMO, 2011). Nevertheless, this dynamical approach is very precise and several studies have benefited from it. Muis et al. (2016) created the Global Tide and Surge Reanalysis (GTSR), reconstructing storm surge levels worldwide based on hydrodynamic modelling by using the Delft3D Flexible Mesh Suite with D-Flow. Vousdoukas et al. (2016) used a similar model, this time forcing the hydrodynamic model with wind and pressure fields from climate models, to study the effect of climate change on extreme storm surge levels along the European coastline. These dynamical approaches produce more homogeneous storm surge historical records, as most of the exploratory data analysis is usually done with tidal gauges, sparse in space and time (Cagigal et al., 2020; Arns et al., 2020; Williams et al., 2016).

Conversely, a number of studies have benefited from the speed of data driven approaches to reconstruct storm surge levels with a reliability similar to that of numerical models, but at a fraction of their computational effort. Using this approach, statistical models can be trained with both observational and numerical model data to produce faster storm surge forecasts. In the field of statistical models, Salmun et al. (2009) and Dangendorf et al. (2014) applied multiple linear regression to model the relationship between surge (as predictand), and wind and sea level pressure (as predictors). Cid et al. (2017) utilized the same linear model with local atmospheric predictors to provide a global storm surge database. Based on this methodology, Cid et al. (2018) reconstructed daily maximum storm surges for the Southeast Asia region, and Cagigal et al. (2020) produced a 0.25º resolution hindcast of storm surge in New Zealand, utilizing this knowledge to obtain storm surge projections until 2,100 with different global climate models. Rueda et al. (2019) used a similar approach to reconstruct both waves and surges in New Zealand (using spatially larger predictors), obtaining values of 0.86 and 0.83 for the pearson correlation coefficient in the Kapiti and Green islands tidal gauges, respectively.

All these studies used linear techniques to reconstruct the storm surge levels, while other studies such as Bruneau et al. (2020) used neural networks for the same purpose, using pre-defined predictors. They obtained promising results for the non-tidal residual (this is the sea level that remains when the astronomical tide is subtracted from the total water level) at a large number of locations over the world, capturing the non-linear relationships between the predictor and the predictand. Very recent studies compare the predictive capabilities of several data driven methods, testing a few atmospheric predictors. For example, Tadesse et al. (2020) showed how slightly different predictors can reconstruct the storm surge around the world, giving a set of potential statistical models. Moreover, Tiggeloven et al. (2021) also evaluated predictor capabilities using neural networks (LSTM, CONV and ConvLSTM) as the main models, obtaining better results as the predictor becomes larger in space and includes more variables such as the wind, or non linear components of the input variables ($u^2$, $v^2$...). In Tiggeloven et al. (2021), probabilistic predictions are obtained, which is a crucial feature of data driven models, and an approach that is impossible with traditional numerical models.

In this work we focus on finding the combination of atmospheric variables that can best represent the behaviour of the storm surge over New Zealand, because although some of the previously mentioned studies try to carry out this process by varying some of the parameters taken into account in this work (Tiggeloven et al. (2021) evaluates different variables with local domains, and Tadesse et al. (2020) also utilizes different variables according to expert knowledge), none of these studies explore the total number of possible combinations within their search space, thus lacking all the necessary results to carry out a certain conclusion. In addition, 3 different statistical models are used in this study to give robustness to the results obtained, and to understand not only which variables can most affect the storm surge, but also the differences in its reconstruction that each of the models can provide. Another important novelty of this methodology is the utilization of the atmospheric conditions from 0 to 72 hours before the storm surge maximum, defined here as time lag. Tadesse et al. (2020) also used lags that covered the previous 30 hours to the observed storm surge maxima, but just some specified statistical models and predictors were tried under this approach, finding the usage of this time lag really improved the performance of the models.

Therefore, this study tests 36 different atmospheric predictors, understanding predictor as the combination of different atmospheric variables that can be used to reconstruct the storm surge signal. The variables used are known to influence the storm

surge, and are the sea-level-pressure fields, the sea-level-pressure fields gradients and the u and v components of the wind, projected to each studied location, to simulate the wind component that contributes to the wind set-up. The spatial and temporal scales used for the predictors are also varied, to understand the effect of different spatial domains and previous atmospheric conditions. To this end, both local and regional predictors are tried, and we test predictors including different time lags (atmospheric conditions existing the previous days to the reconstruction), which is also a crucial novelty of this study. Additionally, 3 different statistical models, including multi-linear regression, $k$-NN regression and gradient boosting regression, are tested, assuring the predictor works for different models. To summarize, in this work we propose a robust methodology that explains why the differences in the results based on different predictors and models might appear, setting a comprehensive framework to find the best performing predictor of daily storm surge maxima and be able of efficiently reconstruct storm surge maximum levels around the islands of New Zealand. Model performance is evaluated with the Kling-Gupta Efficiency (Gupta et al., 2009), as it incorporates 3 sub-metrics: correlation, bias term and variability term, each characterising an important aspect of the prediction performance. This detailed predictor analysis is performed in 29 different locations around New Zealand, where nearby observational data are also available. Once the best performing atmospheric predictor is identified (defined as the atmospheric predictor that best reconstructs the storm surge signal), the reconstruction is extended to the whole of the New Zealand coastline and the results for the 3 statistical models are contrasted.

This article is structured in 6 sections. In Section 2, the databases used are described, then, in Section 3, the methodology is explained in detail. In Section 4, results for all the predictors, models and locations are shown. In Section 5, the main points are summarized in the discussion, where we also address future tasks. Finally, the conclusions are presented in Section 6. The Appendix contains a detailed description for all the statistical models used and the validation metrics for both the numerical model and the best performing predictor and statistical model.

## 2 Databases description

Model predictor data are sourced from a global atmospheric reanalysis. Predictands (storm surge signal) are obtained from a high resolution regional hydrodynamic hindcast for New Zealand waters (Moana v2 hindcast model). In addition, observations acquired from 29 tidal gauges spread around the coast of New Zealand (see Fig.2) are used to validate the sea level data from the hindcast. In Figure 1, the spatial domains of both hindcasts can be seen, where variables are also shown.

### 2.1 Atmospheric data

For the atmospheric data, we use a global reanalysis developed by NCEP (National Centers for Environmental Prediction) in the configuration of CFSR (Climate Forecast System Reanalysis). In order to cover the period of the hydrodynamic data, we use both CFSR (Saha et al., 2010), which extends from 1979 to 2011, and CFSRv2, (Saha et al., 2011), which extends from 2011 to present. We utilize sea-level-pressure (SLP), sea-level-pressure gradients (SLPG) and winds as the main predictors affecting the storm surge. SLP data are used at their native resolution which is 0.4º in both CFSR and CFSRv2. The meridional

and zonal components of the wind, whose original resolution is 0.3º in CFSR and 0.2º in CFSRv2, were interpolated linearly over the same grid as the SLP. All the atmospheric variables used are resampled to daily values calculating the mean.

These wind components are not directly used, they are projected to the location where the reconstruction is made. As illustrated in Figure 1, the wind vectors are projected over the line that joins each point in the atmospheric gridded domain to the location where the storm surge is predicted. Then, if the wind blows to a desired location at time $t$, this wind will contribute positively to the storm surge signal (red arrows). On the other hand, winds blowing in the direction opposite to the desired location will contribute negatively to the storm surge (blue arrows). Finally, land location is also taken into account, so winds blowing directly towards a certain location but from land side are discarded.

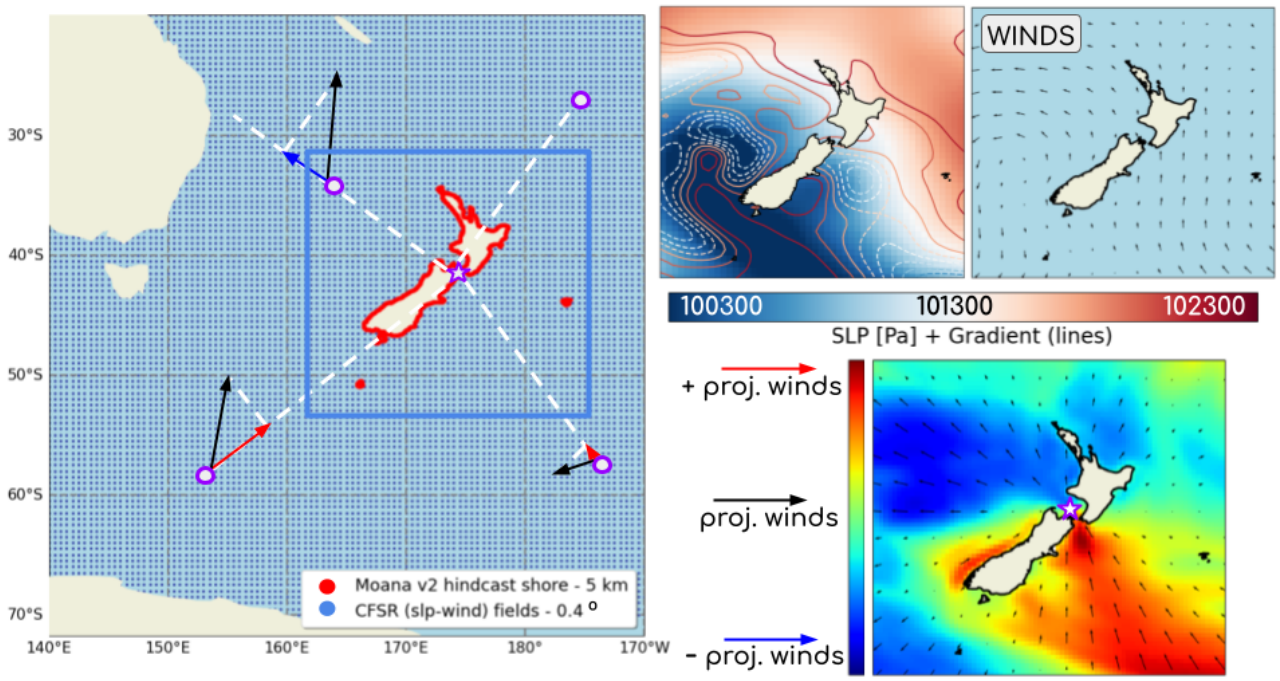


**Figure 1.** The atmospheric reanalysis is represented by blue dots (blue square represents the maximum spatial domain considered in the experiments), the Moana v2 hindcast model output, which corresponds to the storm surge hindcast, is shown in red. The two plots in the top-right represent the sea-level-pressure fields (SLP) and the sea-level-pressure fields gradients (SLPG) (left), and the u10 and v10 components of the wind (right). The arrows in the left panel show how the wind projection is made.

## 2.2 Storm surge data

The storm surge datasets are separated into hindcast and observational data. The hindcast is used here to fit the statistical models due to its spatial and temporal extent, as we want to reconstruct the storm surge maximum levels all over the New Zealand coastline. The observational data, which correspond to tidal gauges, are used to validate this hindcast.

The processing of the total sea level series for both the hindcast and the tidal gauges data was done using the open-source toolbox Toto (https://github.com/calypso-science/Toto). The linear trend was first removed from the time series applying a linear regression to the sea level signal, which is the sea-level-rise in the historical period of the hindcast. Tidal analysis was then carried out using the algorithms implemented in the Python version of the UTide software (Codiga, 2011) and the astronomical tide estimates were used to fill any missing gaps in the tidal gauge data. The monthly mean sea level variation was then removed from the time series using a 30 day rolling mean window. Finally, the storm surge signal was extracted using a Lanczos lowpass filter (Thomson and Emery, 2014), with a cut-off period of 30 hours. Considering that the inertial period around New Zealand latitudes varies from 16h to 22h approximately, the 30 hours cut-off period allowed for both tidal and inertial oscillations to be removed from the total water level, isolating the storm surge signal.

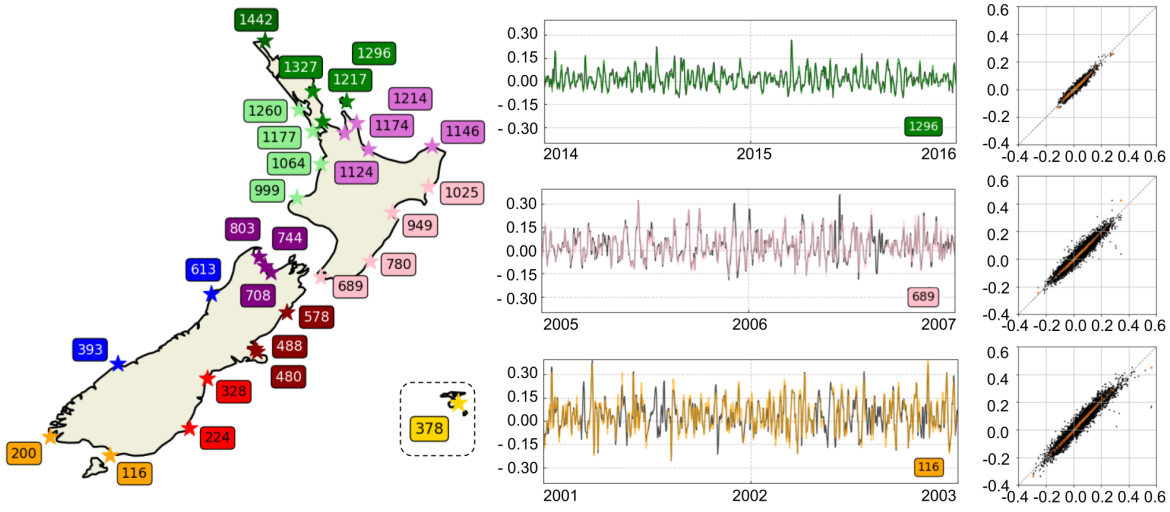


**Figure 2.** Location of tidal gauges along the NZ coast and their closest numerical model nodes (left) and storm surge validation of Moana (numerical data) vs tidal gauges (observational data) at three locations (right). Hindcast is represented in black in the time series plots.

### 2.2.1 Moana v2 hindcast

The numerical model storm surge data along the New Zealand coast are obtained by postprocessing of the sea surface height fields from version 2 of the Moana Backbone Model (Azevedo Correia de Souza et al., 2022) (data are open access and samples are available here (Azevedo Correia de Souza, 2022), while the full dataset can be downloaded from the project webpage at https://www.moanaproject.org/). The Moana Backbone Model is a 25-year (1993 - 2017) regional hydrodynamic hindcast model of New Zealand waters released in 2020 by the New Zealand MetService. The hindcast was produced using the Regional Ocean Modeling System (ROMS), version 3.9, which is a free-surface, terrain-following, hydrostatic numerical model that solves the 3D Reynolds-averaged Navier-Stokes equations using Boussinesq approximation (Haidvogel et al., 2008). The hindcast horizontal resolution is 5 km over the whole domain with 50 levels in the vertical. ROMS was forced with atmospheric conditions from the Climate Forecast System Reanalysis (CFSR) versions 1, (Saha et al., 2010), and 2, (Saha et al., 2011) (reason why we use the same global reanalysis as the predictors to our models). The open boundaries were forced with currents, sea level, temperature and salinity, from the Copernicus Global Ocean Physics Reanalysis (GLORYS) version 12v1 and spectral tidal forcing from the OSU Tidal Inversion software (OTIS) version 7.1.

### 2.2.2 Observational data

We gathered data from 29 tidal gauges located around NZ as shown in Figure 2 and those were used to validate the numerical model. We also used the location of the tidal gauges to select the closest hindcast nodes for which the initial predictor experiments were performed. This was motivated by the fact that the different sub-shores where these tidal gauges are located exhibit different storm surge behaviours. The complexity and varying orientation of New Zealand's coastline mean that there can be strong local differences in storm surge signals, and the tidal gauges are well spread around NZ. Comparisons exhibit a very good correlation between the outputs of the numerical model and the tidal gauges, and thus this storm surge hindcast can be used to calibrate the statistical models (metrics such as the RMSE and the pearson or spearman correlations can be found, for all the validated locations, in Table A1).

## 3 Methodology

The methodology is divided into two steps. We first find the optimal atmospheric predictor (defined as the atmospheric predictor that best reconstructs the storm surge signal), and then use this predictor to reconstruct the historical storm surge maximum levels along the entire coast of New Zealand.

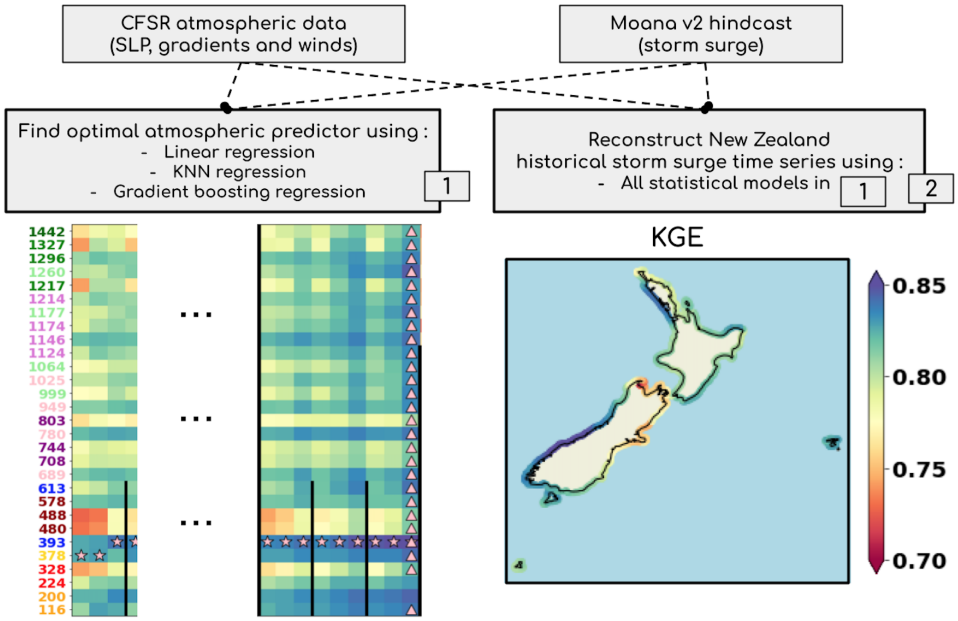


**Figure 3.** A diagram of the analysis workflow with two steps: 1) Find optimal atmospheric predictor (the columns correspond to different models, and the colored numbers as rows represent the locations where the experiments were tested, see Fig.2). 2) Reconstruct the historical storm surge maximum levels along the entire coast of New Zealand.

### 3.1 Find optimal atmospheric predictor

We first identify the optimal atmospheric predictor from a total of more than 3,000 experiments, which include any possible combination of atmospheric predictor, location and statistical model. Specifically, with 36 atmospheric predictors (Table 1), 29 locations to study (Figure 2) and 3 statistical models to evaluate (multi-linear, $k$-NN and gradient boosting regression), this makes a total of 3,132 experiments.

The same workflow is followed by all linear methods as depicted on Figure 4 and involves: 1) predictor building, 2) a dimensionality reduction step using Principal Components Analysis (PCA) (Gutiérrez et al., 2004; Wilks, 2005), 3) fitting the statistical models against the PCA-projected predictor and 4) model evaluation. The four steps are explained in detail below as they summarize the way the optimal atmospheric predictor is found.

#### 3.1.1 Predictor building

The search for the optimal atmospheric predictor is one of the main goals of this work, and hence is why all relevant combinations have been considered. This combinatorial cloud covers all the possible cases summarised in Table 1. The sea-level-pressure fields are always used, and then the gradients and the projected winds are added in turn. For all these variables, past time frames can be used (defined as "time lag" approach in this study), so the information from times preceding the reconstruction are used. Finally, regarding the spatial extent of the predictor, 3 different regions are tested, two local squared regions

of 3·3º and 5·5º centered in the location of interest and a bigger region that encompass the whole area of New Zealand (blue square in Fig.1).

**Table 1.** Variables and parameters used to construct the predictors are shown. Notice the SLP fields are always used, but all the other features might change.

| 4x Data sources | | | 3x Time lag | 3x Region |
|---|---|---|---|---|
| sea-level-pressure fields (SLP) | gradient fields calculated from the SLP variations | projected winds (calculated from u10 and v10) | whether to add previous time steps to reconstruct | spatial region to consider around the location of interest |
| predictor might use:<br>SLP<br>SLP + gradients<br>SLP + projected winds<br>SLP + gradients + projected winds | | | 1 (just time t) | local - 3º x 3º |
| | | | 2 (t and t-1) | local - 5º x 5º |
| | | | 3 (t, t-1 and t-2) | Regional (160,185,-52,-30) |

### 3.1.2 Principal Components Analysis

After the initial predictor matrix is assembled by concatenating the raw predictor data from the atmospheric reanalysis and in order to reduce the number of features fed to the statistical models, a dimensionality reduction step is applied to the predictor matrix. We use PCA (Gutiérrez et al., 2004; Hastie et al., 2001), and retain the leading components ensuring that 98% of the variance is explained.

The predictor is projected in a new space, where the first coordinates in this new space explain the highest percentage of the variance in the data. This data transformation is just an orthogonal linear transformation that converts the data to a new coordinate system such that the greatest variance by some scalar projection of the data is captured by the first coordinate (called the first principal component, $PC_1$), the second greatest variance on the second coordinate, $PC_2$, and so on... where the new weights that will transform the original atmospheric predictor into the new basis are represented in the equation below:

$$\mathbf{w} = \frac{\mathbf{w}\mathbf{X}^T\mathbf{X}\mathbf{w}}{\mathbf{w}^T\mathbf{w}} \rightarrow w_{(1)} = \arg\max\left\{\frac{\mathbf{w}\mathbf{X}^T\mathbf{X}\mathbf{w}}{\mathbf{w}^T\mathbf{w}}\right\} \tag{1}$$

where $\mathbf{X}^T\mathbf{X}$ is the covariance matrix of the original atmospheric data (bold variables imply they are matrices). The relationship between the original and the projected data can be written as: $\mathbf{X}(x,t_i) = \mathbf{EOF}_1(x) \times \mathrm{PC}_1(t_i) + \mathbf{EOF}_2(x) \times \mathrm{PC}_2(t_i) + ... + \mathbf{EOF}_n(x) \times \mathrm{PC}_n(t_i)$, where the PCs represent the contribution of each EOF in time, and the EOFs represent the oscillation modes that can be added together to reconstruct the original timeseries. With this PCA, the data is reprojected in a easily reducible new space, where the first variables in this new space have two principal characteristics: they represent the higher percentage of the variance in the original data (1) and they have physical meaning (2) (Camus et al., 2014).

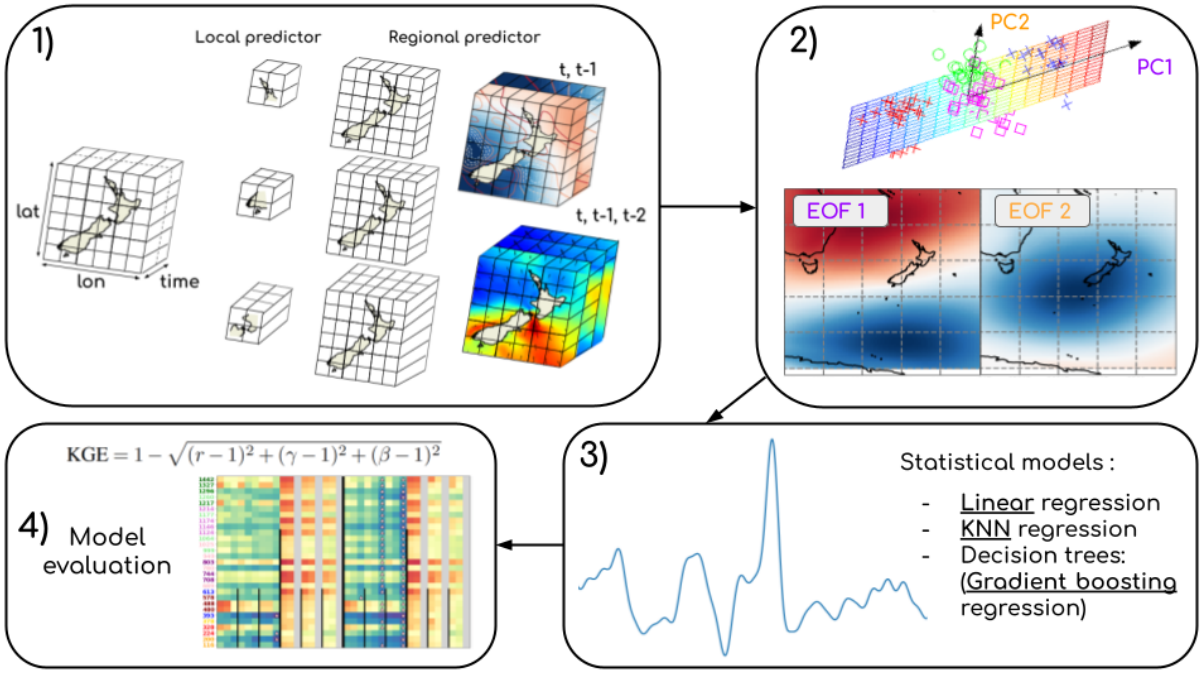


**Figure 4.** Steps for finding the optimal atmospheric predictor. 1) Predictor building, 2) Principal Components Analysis, 3) Statistical models and 4) Model evaluation.

### 3.1.3 Statistical models

In this study, we use the 70% of the data for training, and the remaining 30% are used to validate the models. The training data use the first 17 years of the numerical model, and the remaining 8 years are used to validate. Here, we are considering that the general behaviour of the atmosphere does not change between the training and testing periods that we are using. Different training and validations periods, even shuffling the data, were tested. Best performance of the statistical models, in terms of KGE metric, was obtained using at least 70% of data for training.

We use linear regression, $k$-NN regression and gradient boosting regression algorithms to obtain the best possible atmospheric predictor from all the candidates. A brief explanation of these models is given below, while a more detailed one can be found in the Appendix.

1. **Linear regression**: The method fits the hyperplane (a hyper-dimensional plane with dimensions equal to the number of PCs used, $N$) to the data so the squared errors between the predicted and the real values are minimized. The optimal linear regression parameters are given by $\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$.

2. **$k$-NN regression**: Given a dataset, each time a new prediction is made, the model predicts the mean of the target values of the $k$-nearest neighbors based on the Euclidean distance (Altman, 1992). In this study, we experiment with number of

neighbors varying from 1 to 31 in steps of 3, so 11 $k$-NN models are calibrated for each combination of predictor and location.

3. **Gradient boosting regression**: The method combines the ability of weak decision trees to obtain robust predictions of a target variable, optimizing the tree's individual ability by minimizing a loss function, calculating its gradient at each step (Friedman, 2000). In this case, we change both the maximum depth of the tree, testing 3, 9 and 15 final nodes, and the minimum percentage of the data available at each split, testing 3, 9 and 15% (trying to see how overfitting might affect the results). Then, 9 gradient boosting models are trained for each combination of predictor and location.

#### 3.1.4 Model evaluation

After running more than 3,000 experiments, we select the Kling-Gupta Efficiency statistic (dimensionless) (Gupta et al., 2009) as the single metric to evaluate all the results, which is defined as:

$$\text{KGE} = 1 - \sqrt{(r-1)^2 + (\gamma-1)^2 + (\beta-1)^2} \tag{2}$$

where $r = \frac{\sum_{i=1}^{p}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{p}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{p}(y_i-\bar{y})^2}}$, $\gamma = \frac{\sigma_{rec}}{\sigma_{obs}}$ and $\beta = \frac{\mu_{rec}}{\mu_{obs}}$, being $r$ the pearson correlation coefficient between reconstructed and observed storm surge data (dimensionless), $\gamma$ the variability ratio (dimensionless), $\beta$ the bias ratio (dimensionless), $\sigma$ the standard deviation, $\mu$ the mean, and the indices $rec$ and $obs$ represent reconstructed and observed storm surge values, respectively. KGE, $r$, $\beta$ and $\gamma$ have their optimum at unity (1), and KGE values bigger than -0.41 ensure the models are performing better than the mean flow prediction (Knoben et al., 2019).

KGE is particularly applicable here as it accounts for three different important aspects of the reconstructed time series by including a correlation, a bias and a variability term, and thus providing information on how well the storm surge time-series are reproduced. For a full discussion of the KGE-statistic and its advantages over the Nash–Sutcliffe efficiency see (Nash and Sutcliffe, 1970) (or with the mean squared error see (Gupta et al., 2009)), as these are the most commonly used metrics for evaluating hydrological model performance, where the behaviour of the models at the extremes is very important too.

### 3.2 Reconstruction of the storm surge along the New Zealand coastline

Finally, using the best atmospheric predictor identified in the first part of the study, the storm surge is reconstructed over the whole coastal domain of the hindcast data, for the 3 statistical models. The procedure involves training separate models for each of the coastal points following the workflow depicted in Figure 4, using only the best atmospheric predictor. In this part, we use again the 70% of the data to train the models (which correspond to almost 17 years of data), and the remaining 30% are used to validate results.

## 4 Results

Results are analyzed in two different sections. The results of the experiments aimed at identifying the best predictor are first analyzed, using Figures 5 and 6, and then, the results of their extension to the whole coast, with the optimal atmospheric predictor, are presented in Figure 7.

### 4.1 Find optimal atmospheric predictor: Experiments results

The results of the experiments run for the different models and predictors at all selected locations are presented, where we can observe the relationship between the input data and the best performing model. Moreover, the relationships between the different variables used and the reconstructed surge levels, as well as the effect of the temporal (time lag) and spatial domains, are shown here, to asses the optimal atmospheric predictor.

#### 4.1.1 Overview of the experiments

In this section we will describe all the features and variables affecting model performance (winds, time lag, spatial domain...), with the aim of highlighting differences between the models.

The main difference we observe in Figure 5 is that predictors using the projected winds are outperformed by those using the SLPG, providing inferior KGE values for almost all the scenarios, and for the 3 statistical models (here we compare the region in the middle of the plots, so Gradients=False and Winds=True, with Gradients=True and Winds=False). Note that some combinations are missing in Fig.5 (shown as grey), and these correspond to the combination of wind and regional predictors; because we are considering the winds are only relevant in the local scale (WMO, 2011). All these scenarios were outperformed by the gradients, which appear more smooth, possibly having more direct information in the storm surge behaviour. When both the SLPG and the projected winds are used, model performance also worsens, compared to just utilizing the SLPG reconstructions.

Also in Figure 5, the results for the time lag parameter suggest it is best to utilize at least the 2 previous days to the moment of the reconstruction, as results show adding time lags improve the performance of the statistical models in more than 95% of the experiments tried. The regional domain gives the best results using the linear regression model, obtaining the best possible results in all locations. This results for the time lag parameter confirm results in previous studies that suggest the storm surge is a long wave that depends on the atmospheric situation over a large spatial domain (Bell and Goring, 1996), but also in the way these atmospheric conditions develop through time.

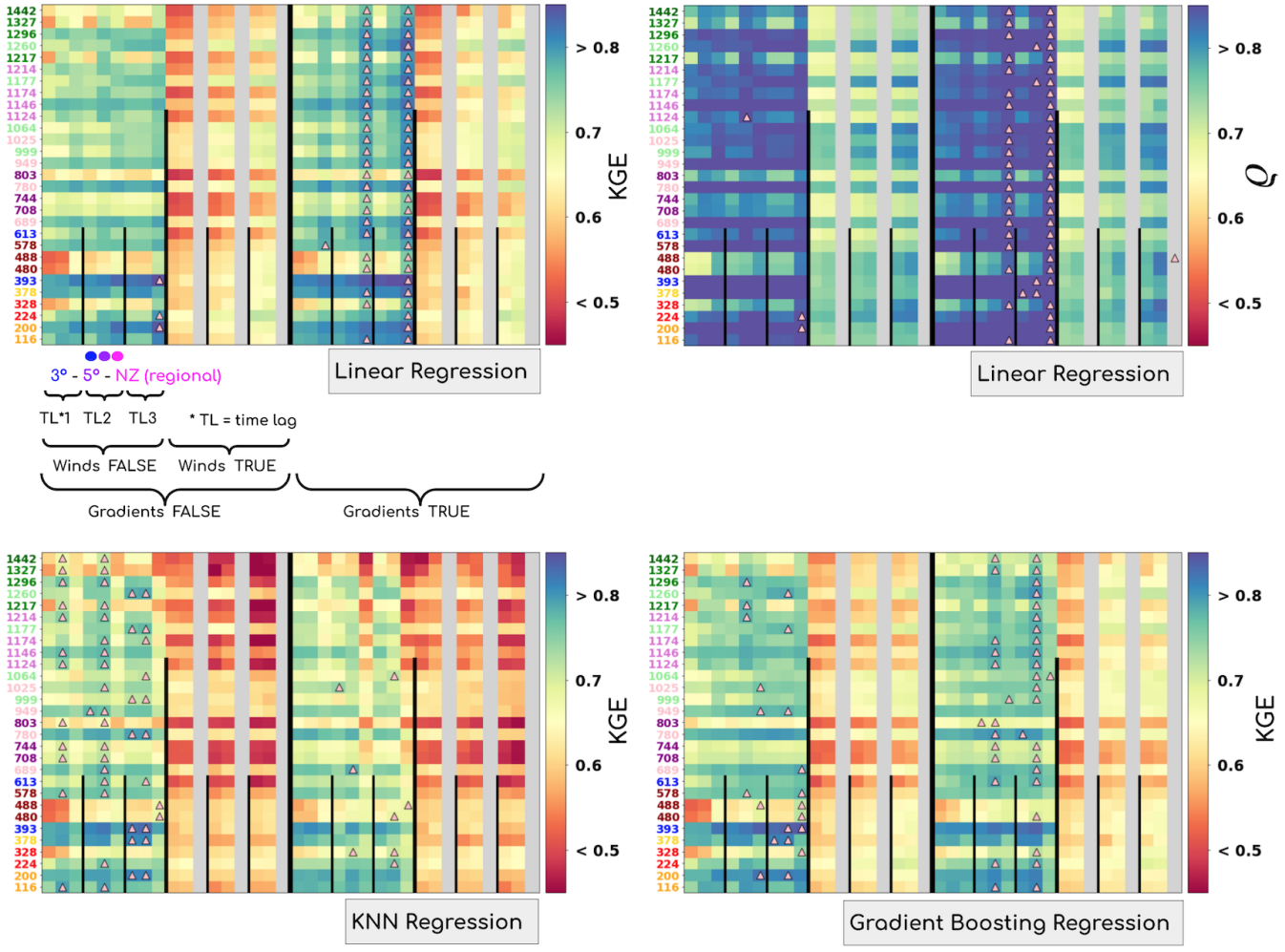


**Figure 5.** Models performance in terms of the Kling-Gupta Efficiency (KGE) and pearson correlation coefficient (only for multi-linear regression) for all the available predictors (x-axis) and in all studied locations (y-axis, colored numbers correspond to the nodes in the numerical model data where the experiments have been tested, see Fig.2), for the three statistical models (multi-linear, $k$-NN and gradient boosting regression). The up triangles appearing in each row represent the best 2 performing predictors.

Models performance is shown in Figure 6, where multi-linear regression provides a better reconstruction as more data are provided as input, achieving the best results when regional predictors and 3 time lags are used, and the reconstruction worsens as we decrease the amount of information fed into the model. Conversely, $k$-NN and gradient boosting regressions, even though it is clear that they also require sufficient information to behave well, perform worse in situations where the regional predictor is used, but work better when local features influencing the storm surge are utilized (outperforming multi-linear regression in these cases.

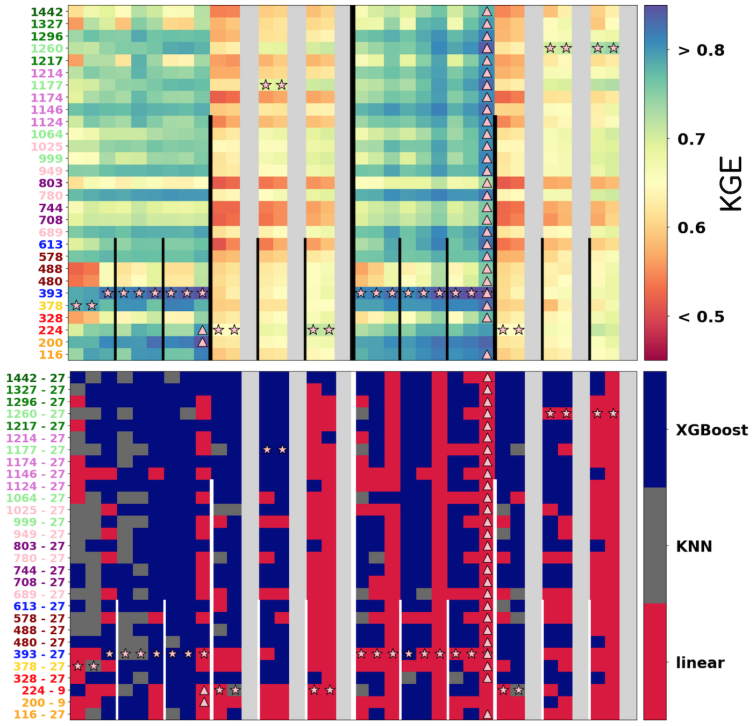


**Figure 6.** top) KGE values for the best statistical model, for all atmospheric predictors. down) Best statistical model, for all the atmospheric predictors. Triangles show the winning predictor for each locations. Stars show the winning location for each predictor.

Now, we focus on the influence of the hyperparameters of the $k$-NN and gradient boosting regressors on model performance, shown in Figure 7. For the gradient boosting model, we tried different tree maximum depths and minimum number of data points per final leaf on the decision trees, and experiments show an optimal convergence at depth≈9 and when at least 3% of the data remain at each final leaf. When the tree is pruned more and leaves start acquiring fewer data points, the model might start overfitting. In the case of $k$-NN, the most informative parameter is the number of neighbors used, which can also help understanding the results presented above. The selection of this hyperparameter is not straightforward and can also lead to overfitting, which is known as the bias-variance trade-off. On the one hand, when the number of neighbors is very low, or even 1, the bias in the reconstructed data is minimal but the overall variance in the data is large. On the other hand, if the model is trained with a large number of neighbors, the bias in the results will become also higher, but the variance will be reduced, and so the capacity to predict extreme events, which in this case can be flooding cases.

### 4.1.2 Best performing predictor selection

The results presented in the previous section do not give us an overall best predictor as predictor performance is highly influenced by the model and the data that is fed into it.

Nevertheless, the best predictor that has been chosen to carry out the reconstruction of the historical storm surge daily maxima along the New Zealand coastline is that associated with the best results provided by the multi-linear regression approach. Indeed as shown in Figure 6, multi-linear regression consistently outperforms the 2 other models and almost always does so using a predictor that includes SLP and SLP gradients, a time lag of 3 days (the day and two days before the reconstruction) and the regional domain. In the two locations where this predictor does not lead to the best results, the best results are obtained when the gradients are not used, but again using multi-linear regression. However, the difference in model performance is below 0.02 for both the KGE and the pearson correlation metrics.

In Figure 8, the daily maximum storm surge time series comparison between numerical model data and multi-linear regression model reconstructions for 5 different locations and 2 different storms are shown. The validation metrics are excellent for this statistical model, which is the best performing model in all the cases. The validation metrics for all the locations studied in the experiments (see Fig.2) are shown in Table A2, complementing the information in this figure.

Both the time series and the scatter plots show promising results in all locations, here we are showing examples of both the best (200, 393 and 1146) and worst (480 and 803) reconstructed locations, from the locations we used in the experiments. Metrics for the mean values show pearson correlation coefficients above 0.85, and KGE values above 0.73. For the extreme metrics, that can be found in Table A2, the RMSE calculated for the 99% percentile show values always below 10cm, reaching almost 4cm errors in the best reconstructed locations.

### 4.2 Reconstruct the storm surge along the New Zealand coastline: Final results

Once the best predictor has been identified, for which the SLP and the gradients, time lapse of 3 and the regional predictor are used, this predictor is used in all the models to reconstruct the historical storm surge daily maximum values in the entire coast of New Zealand. Results for the KGE metric are shown in Figure 7.

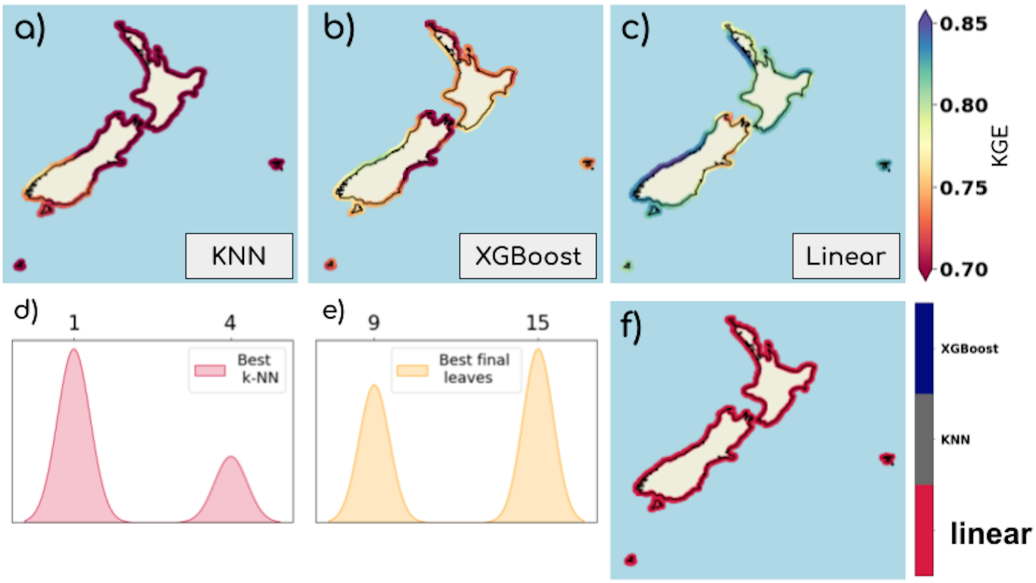


**Figure 7.** Storm surge spatial reconstruction along the New Zealand coastline. Model performance expressed as Kling-Gupta Efficiency (KGE) using $k$-NN regression (a), gradient boosting regression (b) and linear regression (c). Kernel density estimation (KDE) plots of the hyperparameters that provide the best results for: d) the number of neighbors for the $k$-NN regression; e) the leaves for the gradient boosting. Best statistical model for each hindcast node (f).

In agreement with the results presented in the earlier sections, performance is not uniform over the coastline, the regions which are more exposed to storms coming from the south of Tasmania are more predictable in terms of the storm surge maxima and present better results, while regions which are hidden from this dominant storms are more difficult to predict.

The better performance of the multi-linear regression technique was explained before in the experiments results, and we can conclude that highly informative atmospheric predictors work better with multi-linear regression, except from the utilization of the projected winds. However, we can see how $k$-NN and gradient boosting underperform, although gradient boosting provides similar results to its principal competitor, the multi-linear regression. In the case of $k$-NN, where the best results are obtained with a very low number of neighbors, results are always below the other statistical methods, and although extremes are well predicted given this low number of closest neighbors, the overall performance is very poor.

Figure 7 shows the value of the hyperparameters that provide the best performance of the statistical models. For the $k$-NN case, a very low number of neighbors is usually chosen as best by the KGE metric (which is very influenced by the extremes reconstruction), and for the gradient boosting regression, the KDE (Kernel density estimation) has its high at a maximum depth of 15, which represents the more complex tree.

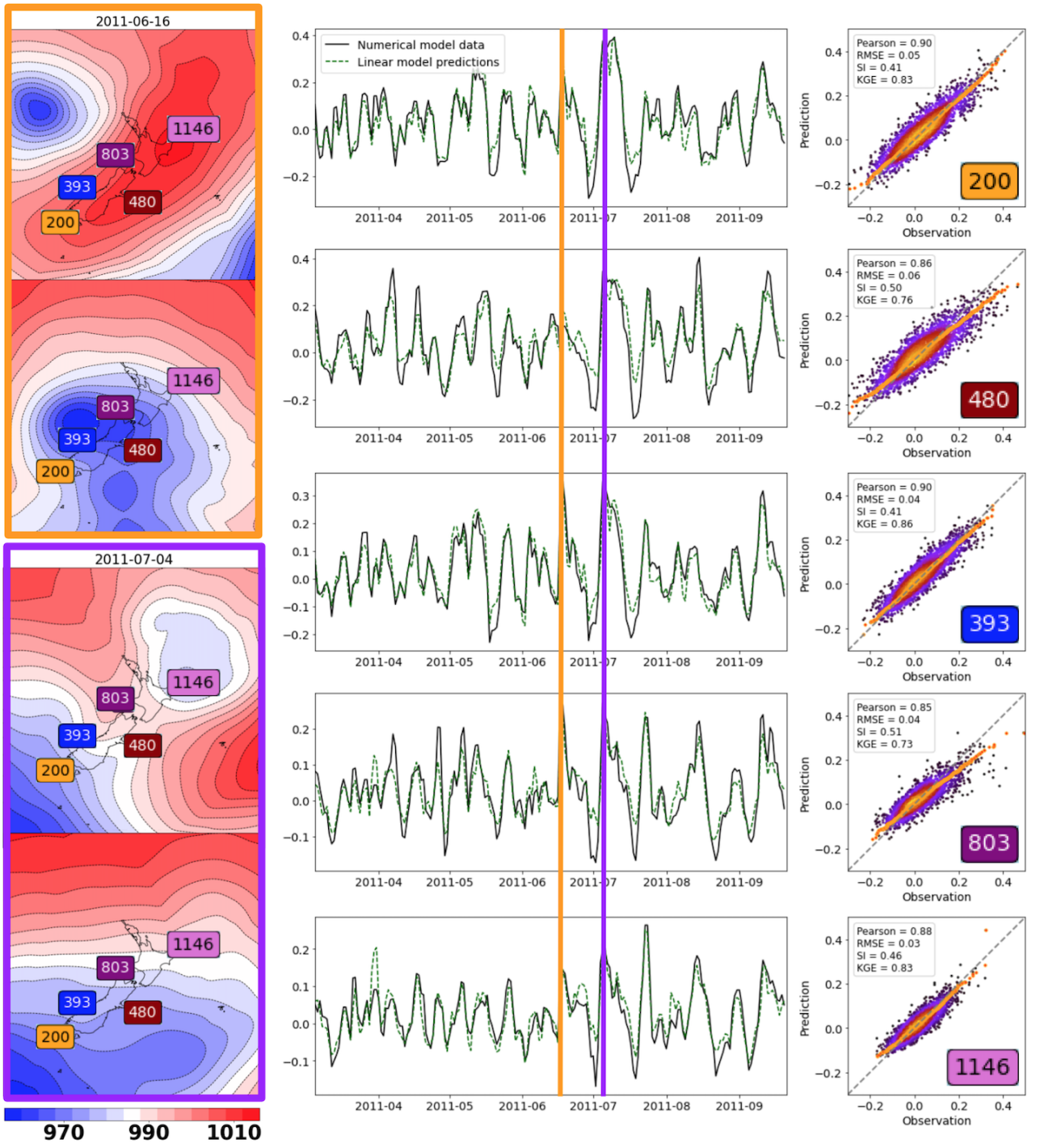


**Figure 8.** Daily maximum surge time series comparison between numerical model data and linear model reconstructions (middle), scatter plots and q-q plots for model data and reconstructions (right) for the validation period (2010-2019) for locations marked in SLP panels (left). SLP panels show the SLP fields for two times belonging to two different storms approaching the islands of New Zealand, represented by colored lines in the time series plots.

## 5 Discussion and future work

We have studied the relationships between several combinations of atmospheric predictor, statistical model and location to reconstruct storm surge along the New Zealand coastline. If we call each combination an experiment, we have evaluated a total of more than 3,000 experiments (many more if we count the repeated experiments depending on the different hyperparameters used in the models), in order to reach the conclusions outlined below.

In addition, the behaviour of all the statistical models was evaluated and the best predictor was used to determine the predictability of the maximum value of storm surge, resolved by location, in the two islands of New Zealand, the North Island (or Te Ika-a-Maui), and the South Island (or Te Waipounamu). Thus, given this best predictor, multi-linear regression is the model that achieves the highest values for our metric of choice, which is the Kling-Gupta Efficiency statistic.

The following are the main conclusions drawn from this study:

### 5.1 Predictors

- The main feature observed in relation to the atmospheric predictor is the systematically positive impact of the SLP gradients on the reconstruction performance. On the contrary, in the case of the local predictors, the projected wind always degrade the performance of all models. This drop in performance might be related to the fact that first PCs representing the 98% of the variance are not that informative in the latter case, adding noise to the reconstructions. To validate the usage of the projected winds, we tested some experiments, using just the wind speed, calculated from the meridional and zonal components, leading to similar results. Besides, wind speed PCs were very similar to the PCs calculated from the SLP gradients, then suggesting the direct comparison of these used variables affecting models performance.

- It has been also demonstrated here that the storm surge can be influenced by the atmospheric conditions existing at least two days before the time of analysis, then assuring the storm surge is not just affected by the atmospheric conditions in the exact moment of the reconstruction, but from previous ones too. The storm surge is a long wave with periods ranging from 2 to several days (WMO, 2011), then affected by the atmospheric conditions existing no just in the exact moment of the analysis. In New Zealand, this is also true, and has been analyzed in several studies (Stephens and Bell; Bell and Goring, 1996).

- In this way, the predictors that have generally achieved better results are the regional ones, which might have atmospheric information that is moving towards the predicted location. Moreover, storm surge is a long wave, and it depends on the atmospheric situation over a large spatial domain (Bell and Goring, 1996), and this regional predictor covers a big area that might see the complete storm behaviour (Stephens et al., 2019b).

### 5.2 Statistical models

Each statistical model behaves differently. Multi-linear regression performs best with the regional predictors when the 2 other models give the best results for when using local predictors. When using local predictor multi-linear regression is almost al-

ways outperformed by either gradient boosting or $k$-NN. The reason for this behaviour might be the large number of principal components that are retained with the daily and regional predictors, see Wilks (2005) for a detailed explanation on how a larger number of independent variables might affect the dependent variable reconstruction. In the case of gradient boosting regression, for example, the selection process to bifurcate the tree becomes very costly and imprecise, and the feature engineering step becomes very important here (Hastie et al., 2001; Friedman, 2000). This is not the case of smaller (local of 3·3 or 5·5 degrees) predictors, in which the local atmospheric behaviours are more represented in the first PCs, then being the use of these techniques very suitable.

### 5.3 Location

The spatial pattern of performance observed is clear for all models: the areas more exposed to storms coming from the south-west present more predictable surges, while those areas more sheltered from the same storms are harder to predict. This is because in the latter case the storm has already interacted with the land by the time it reaches these locations and the surge has interacted with the coastline so it is much more complex to predict. This task is even more challenging in areas with high variations of storm surge, as is generally the South Island (Te Waipounamu). Secondly, the worst reconstructions correspond to large embayment areas (Tasman, Golden Bay and the Firth of Thames), where non-linear phenomenal might contribute to the total surge. Finally, the north area of the North island is more affected by meso-scale eddies whose signal was in the same frequency range as the storm surge and might have an impact on the results.

In this study, we propose the utilization of linear models (or diversions of linear models) to predict the storm surge daily maximum levels in New Zealand, to build an understanding on the atmospheric conditions affecting the storm surge behaviour, so the insight gained in this study can be used to mitigate future sources of uncertainty. Future work will include the training of new statistical models that are able to exploit the non-linear relationships between the predictor and the predictand. For example, Bruneau et al. (2020); Tiggeloven et al. (2021) demonstrated that neural networks are useful to predict storm surge and total water levels worldwide, which are to our knowledge the only other studies that have looked at either daily or hourly surge predictions using neural networks at the global scale. Other studies such as Adeli et al. (2022); Bai and Xu (2022) also studied the usage of CNN and LSTM (convolutional and long-short term memory neural networks) to predict storm surge levels, but the problem was slightly different, as they used the parameterized track to train the models, not the spatial atmospheric conditions. This approach is particularly suited to areas that are primarily influenced by TCs, but less in the case of New Zealand. However, using data driven techniques instead of hydrodynamic models lacks the understanding of the underlying physical processes of surges, then leading to several challenges. In order to solve this problem, physics-informed machine learning techniques show up as a future work that can be used to improve the predictive ability for generalizable NN models (Willard et al., 2020; Kashinath et al., 2021b, a). An example of this methods is the sparse identification of nonlinear dynamical systems (Brunton et al., 2016b, a), where data can be used to extract the underlying functions that govern the behaviour of the predictands, being the final relationships comparable to empirical equations, as storm surge can be quickly inferred from atmospheric data (WMO, 2011). Also in this sense, different meta-models can offer very interesting solutions, as

we can run several case studies that clearly represent the overall storm surge behavior, and then use these cases to reconstruct the total performance of the surges in the desired area (Camus et al., 2013, 2011).

The filtered storm surge signal is reconstructed in this work (we apply a Lanczos filter to remove the high frequency oscillations from the non-tidal residual), using atmospheric variables that we know directly influence this variable and were used in several studies (Cid et al., 2017; Cagigal et al., 2020; Tiggeloven et al., 2021). Nevertheless, on the forecasting of a flooding event, we might be interested not only on the storm surge, but in the non-tidal residual as a whole, since Arns et al. (2020) have found that the linear summation of astronomical tide and storm surge might produce overestimation of the total water level. This non-tidal residuals have information that might be predictable, probably adding more informative variables to the predictor set, or even using the information in the actual series. (Bruneau et al., 2020) reconstructed the non-tidal residuals adding waves and precipitation to the predictor variables, and they found decent results worldwide, but difference is not made to distinguish how well models are capturing the actual high frequency residuals in all the locations studied. Few previous works also studied the skew surge, which is the absolute difference between the max recorded sea level and the predicted tidal high water within a tidal cycle (Williams et al., 2016), which might be enough if just the maximum total water level within a tidal cycle is required. Alternatives to solve this problem include the prediction of the tide-surge residuals individually, as a good estimation of the contribution of this part to the full signal has been found to be very important worldwide ((Stephens and Bell) studied this in New Zealand).

Finally, our models can be also trained to reconstruct not just the current maximum surge levels, but future times, then producing reliable forecasts for the sea level in the future days, or even hours if the resolution in both predictors and predictands is changed (Dullaart et al., 2020). Besides, more accurate atmospheric products from local sources such as remote sensing products, even though spatial and temporal data coverage may be limited, might be used to address this problem.

## 6 Conclusions

A novel methodology is proposed in this article that reconstructs the storm surge maximum levels over the entire coast of New Zealand. First, several atmospheric predictors are utilized that use different variables, various time lags and different spatial domains, given 3 statistical models (multi-linear regression, $k$-NN regression and gradient boosting regression), to find the best possible combination. Following this, the storm surge daily maxima is reconstructed with the different statistical models along the entire coast, based on the best performing predictor. In this way, we have demonstrated that careful selection of the combination of atmospheric predictors and statistical model, can lead to large improvements in the ability to predict storm surges. For the multi-linear regression, that is the best performing method, the conclusion is that the more data is used, except from the projected winds, the better this model works. For the other two methods, the increase of the spatial domain of the data might wrongly affect the performance of some of the models. In all the scenarios, using at least two previous day to the reconstruction improve model performance.

Finally, and given a regional predictor with three days historical memory, the storm surge time series in New Zealand have been reconstructed daily using multi-linear regression. Results show very good performance, with average values of 0.88 for

the pearson correlation coefficient and 4.3cm for the root mean squared error metric (RMSE) (the average value for the RMSE in the 99% percentile is 8.2cm). For the KGE statistic, which is the metric used to rank the models, the average value is 0.82. Our results highlight the suitability of data driven models to simulate storm surge maximum levels, and prove the methodology is appropriate for finding a well performing atmospheric predictor that is capable of reconstructing these values, given different data driven methods.

*Acknowledgements.* The authors would like to acknowledge the National Centers for Environmental Prediction for generating the atmospheric data used in this study, the MetService in New Zealand for producing the storm surge hindcast and the New Zealand councils for providing the tidal gauges data (Waikato Regional council, Port Taranaki, Tasman District Council , Northland, District, Greater Wellington, Land Information New Zealand and University of Hawaii Sea Level Center.). This study is funded by the New Zealand Ministry of Business Innovation and Employment, under contract number MSVC1901.

## References


Adeli, E., Sun, L., Wang, J., and Taflanidis, A. A.: An advanced spatio-temporal convolutional recurrent neural network for storm surge predictions, https://doi.org/10.48550/ARXIV.2204.09501, 2022.

Altman, N. S.: An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression, The American Statistician, 46, 175–185, http://www.jstor.org/stable/2685209, 1992.

Arns, A., Wahl, T., Wolff, C., Vafeidis, A. T., Haigh, I. D., Woodworth, P., Niehüser, S., and Jensen, J.: Non-linear interaction modulates global extreme sea levels, coastal flood exposure, and impacts, Nature Communications, 11, 1918, https://doi.org/10.1038/s41467-020-15752-5, 2020.

Azevedo Correia de Souza, J. M.: Moana Ocean Hindcast, https://doi.org/10.5281/zenodo.5895265, 2022.

Azevedo Correia de Souza, J. M., Suanda, S. H., Couto, P. P., Smith, R. O., Kerry, C., and Roughan, M.: Moana Ocean Hindcast – a 25+ years simulation for New Zealand Waters using the ROMS v3.9 model, EGUsphere, 2022, 1–34, https://doi.org/10.5194/egusphere-2022-41, 2022.

Bai, L.-H. nd Xu, H. Accurate storm surge forecasting using the encoder–decoder long short term memory recurrent neural network, Physics of Fluids, 34, 016 601, https://doi.org/10.1063/5.0081858, 2022.

Bell, R., Goring, D., and de Lange, W.: Sea-level change and storm surges in the context of climate change, Institution of Professional Engineers New Zealand Transactions, 27, 1–10, 2000.

Bell, R. G. and Goring, D. G.: Techniques for analyzing sea level records around New Zealand, Marine Geodesy, 19, 77–98, https://doi.org/10.1080/01490419609388071, 1996.

Brenstrum, E.: The cyclone of 1936: the most destructive storm of the Twentieth Century?, Weather and Climate, 20, 23–27, 2000.

Bruneau, N., Polton, J., Williams, J., and Holt, J.: Estimation of global coastal sea level extremes using neural networks, Environmental Research Letters, 15, 074 030, https://doi.org/10.1088/1748-9326/ab89d6, 2020.

Brunton, S. L., Proctor, J. L., and Kutz, J. N.: Sparse Identification of Nonlinear Dynamics with Control (SINDYc)**SLB acknowledges support from the U.S. Air Force Center of Excellence on Nature Inspired Flight Technologies and Ideas (FA9550-14-1-0398). JLP thanks Bill and Melinda Gates for their active support of the Institute of Disease Modeling and their sponsorship through the Global Good Fund. JNK acknowledges support from the U.S. Air Force Office of Scientific Research (FA9550-09-0174)., IFAC-PapersOnLine, 49, 710–715, https://doi.org/https://doi.org/10.1016/j.ifacol.2016.10.249, 10th IFAC Symposium on Nonlinear Control Systems NOLCOS 2016, 2016a.

Brunton, S. L., Proctor, J. L., and Kutz, J. N.: Discovering governing equations from data by sparse identification of nonlinear dynamical systems, Proceedings of the National Academy of Sciences, 113, 3932–3937, https://doi.org/10.1073/pnas.1517384113, 2016b.

Cagigal, L., Rueda, A., Castanedo, S., Cid, A., Perez, J., Stephens, S. A., Coco, G., and Méndez, F. J.: Historical and future storm surge around New Zealand: From the 19th century to the end of the 21st century, International Journal of Climatology, 40, 1512–1525, https://doi.org/https://doi.org/10.1002/joc.6283, 2020.

Camus, P., Mendez, F. J., and Medina, R.: A hybrid efficient method to downscale wave climate to coastal areas, Coastal Engineering, 58, 851 – 862, https://doi.org/https://doi.org/10.1016/j.coastaleng.2011.05.007, 2011.

Camus, P., Mendez, F. J., Medina, R., Tomas, A., and Izaguirre, C.: High resolution downscaled ocean waves (DOW) reanalysis in coastal areas, Coastal Engineering, 72, 56 – 68, https://doi.org/https://doi.org/10.1016/j.coastaleng.2012.09.002, 2013.

Camus, P., Méndez, F. J., Losada, I. J., Menéndez, M., Espejo, A., Pérez, J., Rueda, A., and Guanche, Y.: A method for finding the optimal predictor indices for local wave climate conditions, Ocean Dynamics, 64, 1025–1038, https://doi.org/https://doi.org/10.1007/s10236-014-0737-2, 2014.

Cid, A., Camus, P., Castanedo, S., Méndez, F. J., and Medina, R.: Global reconstructed daily surge levels from the 20th Century Reanalysis (1871–2010), Global and Planetary Change, 148, 9–21, https://doi.org/https://doi.org/10.1016/j.gloplacha.2016.11.006, 2017.

Cid, A., Wahl, T., Chambers, D. P., and Muis, S.: Storm Surge Reconstruction and Return Water Level Estimation in Southeast Asia for the 20th Century, Journal of Geophysical Research: Oceans, 123, 437–451, https://doi.org/https://doi.org/10.1002/2017JC013143, 2018.

Codiga, D.: Unified tidal analysis and prediction using the UTide Matlab functions, https://doi.org/10.13140/RG.2.1.3761.2008, 2011.

Dangendorf, S., Müller-Navarra, S., Jensen, J., Schenk, F., Wahl, T., and Weisse, R.: North Sea Storminess from a Novel Storm Surge Record since AD 1843, Journal of Climate, 27, 3582–3595, http://www.jstor.org/stable/26193434, 2014.

de Lange, W. and Gibb, J.: Seasonal, interannual, and decadal variability of storm surges at Tauranga, New Zealand, New Zealand Journal of Marine and Freshwater Research - N Z J MAR FRESHWATER RES, 34, 419–434, https://doi.org/10.1080/00288330.2000.9516945, 2000.

Dullaart, J. C. M., Muis, S., Bloemendaal, N., and Aerts, J. C. J. H.: Advancing global storm surge modelling using the new ERA5 climate reanalysis, Climate Dynamics, 54, 1007–1021, https://doi.org/10.1007/s00382-019-05044-0, 2020.

Fix, E. and Hodges, J. L.: Discriminatory Analysis - Nonparametric Discrimination: Consistency Properties, International Statistical Review, 57, 238, 1989.

Friedman, J.: Greedy Function Approximation: A Gradient Boosting Machine, The Annals of Statistics, 29, https://doi.org/10.1214/aos/1013203451, 2000.

Goring, D. G. and Bell, R. G.: Distilling information from patchy tide gauge records: The New Zealand experience, Marine Geodesy, 19, 63–76, https://doi.org/10.1080/01490419609388070, 1996.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, Journal of Hydrology, 377, 80–91, https://doi.org/https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Gutiérrez, J., Ancell, R., Cofiño, A., and Sordo, C.: Redes Neuronales y Probabilísticas en las Ciencias Atmosféricas, 2004.

Haidvogel, D., Arango, H., Budgell, W., Cornuelle, B., Curchitser, E., Di Lorenzo, E., Fennel, K., Geyer, W., Hermann, A., Lanerolle, L., Levin, J., McWilliams, J., Miller, A., Moore, A., Powell, T., Shchepetkin, A., Sherwood, C., Signell, R., Warner, J., and Wilkin, J.: Ocean forecasting in terrain-following coordinates: Formulation and skill assessment of the Regional Ocean Modeling System, Journal of Computational Physics, 227, 3595–3624, https://doi.org/https://doi.org/10.1016/j.jcp.2007.06.016, predicting weather, climate and extreme events, 2008.

Hastie, T., Tibshirani, R., and Friedman, J.: The Elements of Statistical Learning, Springer Series in Statistics, Springer New York Inc., New York, NY, USA, 2001.

Heath, R. A.: Significance of storm surges on the New Zealand coast, New Zealand Journal of Geology and Geophysics, 22, 259–266, https://doi.org/10.1080/00288306.1979.10424224, 1979.

Kashinath, K., Mustafa, M., Albert, A., Wu, J.-L., Jiang, C., Esmaeilzadeh, S., Azizzadenesheli, K., Wang, R., Chattopadhyay, A., Singh, A., Manepalli, A., Chirila, D., Yu, R., Walters, R., White, B., Xiao, H., Tchelepi, H. A., Marcus, P., Anandkumar, A., Hassanzadeh, P., and Prabhat, n.: Physics-informed machine learning: case studies for weather and climate modelling, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 379, 20200 093, https://doi.org/10.1098/rsta.2020.0093, 2021a.

Kashinath, K., Mustafa, M., Albert, A., Wu, J.-L., Jiang, C., Esmaeilzadeh, S., Azizzadenesheli, K., Wang, R., Chattopadhyay, A., Singh, A., Manepalli, A., Chirila, D., Yu, R., Walters, R., White, B., Xiao, H., Tchelepi, H. A., Marcus, P., Anandkumar, A., Hassanzadeh, P., and Prabhat, n.: Physics-informed machine learning: case studies for weather and climate modelling, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 379, 20200 093, https://doi.org/10.1098/rsta.2020.0093, 2021b.

Kling, H., Fuchs, M., and Paulin, M.: Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios, Journal of Hydrology, 424-425, 264–277, https://doi.org/https://doi.org/10.1016/j.jhydrol.2012.01.011, 2012.

Knoben, W., Freer, J., and Woods, R.: Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores, Hydrology and Earth System Sciences, 23, 4323–4331, https://doi.org/10.5194/hess-23-4323-2019, 2019.

Muis, S., Verlaan, M., Winsemius, H. C., Aerts, J. C. J. H., and Ward, P. J.: A global reanalysis of storm surges and extreme sea levels, Nature Communications, 7, 11 969, https://doi.org/10.1038/ncomms11969, 2016.

Nash, J. and Sutcliffe, J.: River flow forecasting through conceptual models part I — A discussion of principles, Journal of Hydrology, 10, 282–290, https://doi.org/https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

Rueda, A., Cagigal, L., Antolínez, J. A. A., Albuquerque, J. C., Castanedo, S., Coco, G., and Méndez, F. J.: Marine climate variability based on weather patterns for a complicated island setting: The New Zealand case, International Journal of Climatology, 39, 1777–1786, https://doi.org/https://doi.org/10.1002/joc.5912, 2019.

Saha, S., Moorthi, S., Pan, H.-L., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J., Behringer, D., Liu, H., Stokes, D., Grumbine, R., Gayno, G., Wang, J., Hou, Y.-T., Chuang, H.-Y., Juang, H.-M. H., Sela, J., Iredell, M., Treadon, R., Kleist, D., Delst, P. V., Keyser, D., Derber, J., Ek, M., Meng, J., Wei, H., Yang, R., Lord, S., van den Dool, H., Kumar, A., Wang, W., Long, C., Chelliah, M., Xue, Y., Huang, B., Schemm, J.-K., Ebisuzaki, W., Lin, R., Xie, P., Chen, M., Zhou, S., Higgins, W., Zou, C.-Z., Liu, Q., Chen, Y., Han, Y., Cucurull, L., Reynolds, R. W., Rutledge, G., and Goldberg, M.: NCEP Climate Forecast System Reanalysis (CFSR) Selected Hourly Time-Series Products, January 1979 to December 2010, https://doi.org/10.5065/D6513W89, 2010.

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Behringer, D., Hou, Y.-T., ya Chuang, H., Iredell, M., Ek, M., Meng, J., Yang, R., Mendez, M. P., van den Dool, H., Zhang, Q., Wang, W., Chen, M., and Becker, E.: NCEP Climate Forecast System Version 2 (CFSv2) Selected Hourly Time-Series Products, https://doi.org/10.5065/D6N877VB, 2011.

Salmun, H., Molod, A., Buonaiuto, F., Wisniewska, K., and Clarke, K.: East Coast Cool-Weather Storms in the New York Metropolitan Region, Journal of Applied Meteorology and Climatology - J APPL METEOROL CLIMATOL, 48, 2320–2330, https://doi.org/10.1175/2009JAMC2183.1, 2009.

Siek, M.: Predicting Storm Surges: Chaos, Computational Intelligence, Data Assimilation and Ensembles: UNESCO-IHE PhD Thesis, CRC Press, 2019.

Stephens and Bell: Toolbox 2.2.2: Causes of sea level variability, pp. 13–15, https://niwa.co.nz/sites/default/files/tool_2.2.2_causes_of_sea_level_variability_0.pdf.

Stephens, S., Coco, G., and Bryan, K.: Numerical Simulations of Wave Setup over Barred Beach Profiles: Implications for Predictability, Journal of Waterway Port Coastal and Ocean Engineering, 137, https://doi.org/10.1061/(ASCE)WW.1943-5460.0000076, 2011.

Stephens, S., Bell, R., and Haigh, I.: Spatial and temporal analysis of extreme sea level and skew surge events around the coastline of New Zealand, https://doi.org/10.5194/nhess-2019-353, 2019a.

Stephens, S., Bell, R., and Haigh, I.: Spatial and temporal analysis of extreme sea level and skew surge events around the coastline of New Zealand, https://doi.org/10.5194/nhess-2019-353, 2019b.

Stephens, S., Bell, R., and Haigh, I.: Spatial and temporal analysis of extreme storm-tide and skew-surge events around the coastline of New Zealand, Natural Hazards and Earth System Sciences, 20, 783–796, https://doi.org/10.5194/nhess-20-783-2020, 2020.

Tadesse, M., Wahl, T., and Cid, A.: Data-Driven Modeling of Global Storm Surges, Frontiers in Marine Science, 7, 260, https://doi.org/10.3389/fmars.2020.00260, 2020.

Thomson, R. E. and Emery, W. J.: Data Analysis Methods in Physical Oceanography, Elsevier, Boston, third edition edn., https://doi.org/https://doi.org/10.1016/B978-0-12-387782-6.05001-8, 2014.

Tiggeloven, T., Couasnon, A., van Straaten, C., Muis, S., and Ward, P. J.: Exploring deep learning capabilities for surge predictions in coastal areas, Scientific Reports, 11, 17 224, https://doi.org/10.1038/s41598-021-96674-0, 2021.

Vousdoukas, M. I., Voukouvalas, E., Annunziato, A., Giardino, A., and Feyen, L.: Projections of extreme storm surge levels along Europe, Climate Dynamics, 47, 3171–3190, https://doi.org/10.1007/s00382-016-3019-5, 2016.

Wang, X., Swail, V., and Cox, A.: Dynamical versus statistical downscaling methods for ocean wave heights, International Journal of Climatology, 30, 317 – 332, https://doi.org/10.1002/joc.1899, 2009.

Wikipedia: Gauss-Markov theorem, https://en.wikipedia.org/wiki/Gauss%E2%80%93Markov_theorem.

Wilks, D.: Statistical Methods in the Atmospheric Sciences, Volume 91, Second Edition (International Geophysics), 2005.

Willard, J., Jia, X., Xu, S., Steinbach, M., and Kumar, V.: Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems, https://doi.org/10.48550/ARXIV.2003.04919, 2020.

Williams, J., Horsburgh, K. J., Williams, J. A., and Proctor, R. N. F.: Tide and skew surge independence: New insights for flood risk, Geophysical Research Letters, 43, 6410–6417, https://doi.org/https://doi.org/10.1002/2016GL069522, 2016.

WMO: World Meteorological Organization (2011) Guide to storm surge forecasting, WMO-No 1076, 120pp, https://doi.org/10.25607/OBP-1514, 2011.

## Appendix A: Validation metrics

**Table A1.** Moana Backbone Model (Azevedo Correia de Souza et al., 2022) validation against tidal gauges. The numbers in the left column correspond to the nodes in the numerical dataset and data is hourly validated.

| | BIAS | RMSE | RMSE_99 | SI | EXPL_VAR | PEARSON | SPEARMAN |
|---|---|---|---|---|---|---|---|
| 116 | -0.0 | 0.045 | 0.121 | 0.404 | 0.837 | 0.915 | 0.915 |
| 200 | -0.0 | 0.031 | 0.047 | 0.484 | 0.765 | 0.876 | 0.867 |
| 224 | -0.0 | 0.037 | 0.07 | 0.397 | 0.842 | 0.919 | 0.916 |
| 328 | -0.0 | 0.036 | 0.066 | 0.487 | 0.763 | 0.878 | 0.87 |
| 378 | 0.0 | 0.035 | 0.045 | 0.538 | 0.711 | 0.853 | 0.843 |
| 393 | 0.0 | 0.033 | 0.07 | 0.386 | 0.85 | 0.922 | 0.915 |
| 480 | -0.0 | 0.029 | 0.057 | 0.453 | 0.795 | 0.893 | 0.88 |
| 488 | -0.0 | 0.036 | 0.055 | 0.395 | 0.844 | 0.919 | 0.91 |
| 578 | 0.0 | 0.025 | 0.042 | 0.415 | 0.828 | 0.91 | 0.897 |
| 613 | -0.001 | 0.038 | 0.054 | 0.45 | 0.797 | 0.895 | 0.892 |
| 689 | 0.0 | 0.044 | 0.087 | 0.443 | 0.8 | 0.902 | 0.898 |
| 708 | -0.0 | 0.037 | 0.074 | 0.405 | 0.836 | 0.915 | 0.908 |
| 744 | -0.0 | 0.048 | 0.096 | 0.522 | 0.728 | 0.893 | 0.888 |
| 780 | -0.0 | 0.034 | 0.082 | 0.525 | 0.724 | 0.857 | 0.837 |
| 803 | -0.0 | 0.032 | 0.067 | 0.489 | 0.761 | 0.873 | 0.865 |
| 949 | 0.0 | 0.041 | 0.093 | 0.385 | 0.852 | 0.924 | 0.92 |
| 999 | -0.0 | 0.031 | 0.066 | 0.477 | 0.772 | 0.885 | 0.869 |
| 1025 | -0.0 | 0.04 | 0.075 | 0.395 | 0.844 | 0.926 | 0.923 |
| 1064 | -0.0 | 0.034 | 0.067 | 0.454 | 0.794 | 0.892 | 0.883 |
| 1124 | 0.0 | 0.04 | 0.063 | 0.44 | 0.805 | 0.911 | 0.904 |
| 1146 | 0.001 | 0.029 | 0.038 | 0.484 | 0.765 | 0.877 | 0.848 |
| 1174 | -0.0 | 0.059 | 0.103 | 0.729 | 0.469 | 0.765 | 0.761 |
| 1177 | -0.0 | 0.044 | 0.052 | 0.547 | 0.7 | 0.845 | 0.858 |
| 1214 | -0.0 | 0.036 | 0.092 | 0.388 | 0.849 | 0.922 | 0.915 |
| 1217 | -0.0 | 0.041 | 0.099 | 0.456 | 0.788 | 0.887 | 0.884 |
| 1260 | 0.0 | 0.085 | 0.269 | 1.283 | -0.645 | 0.372 | 0.401 |
| 1296 | -0.002 | 0.059 | 0.074 | 0.682 | 0.534 | 0.804 | 0.781 |
| 1327 | -0.0 | 0.043 | 0.089 | 0.467 | 0.782 | 0.901 | 0.892 |
| 1442 | 0.0 | 0.049 | 0.055 | 0.57 | 0.675 | 0.841 | 0.852 |

**Table A2.** Best experiment metrics are shown, for the best performing atmospheric predictor and using multi-linear regression. Metrics are shown for the locations tested in the experiments.

| | BIAS | RMSE | RMSE_99 | SI | EXPL_VAR | PEARSON | SPEARMAN | KGE | $KGE_{\beta}$ | $KGE_{\gamma}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 116 | -0.005 | 0.054 | 0.102 | 0.418 | 0.8 | 0.895 | 0.887 | 0.821 | 1.101 | 0.897 |
| 200 | -0.004 | 0.046 | 0.068 | 0.409 | 0.812 | 0.901 | 0.893 | 0.834 | 1.098 | 0.909 |
| 224 | -0.004 | 0.051 | 0.072 | 0.449 | 0.78 | 0.883 | 0.877 | 0.797 | 1.112 | 0.877 |
| 328 | -0.003 | 0.057 | 0.137 | 0.476 | 0.752 | 0.868 | 0.864 | 0.773 | 1.08 | 0.834 |
| 378 | -0.001 | 0.043 | 0.083 | 0.428 | 0.796 | 0.892 | 0.888 | 0.828 | 1.02 | 0.867 |
| 393 | -0.002 | 0.04 | 0.07 | 0.409 | 0.814 | 0.903 | 0.897 | 0.857 | 1.074 | 0.925 |
| 480 | -0.002 | 0.058 | 0.123 | 0.495 | 0.732 | 0.857 | 0.848 | 0.756 | 1.07 | 0.815 |
| 488 | -0.002 | 0.058 | 0.126 | 0.496 | 0.731 | 0.856 | 0.847 | 0.755 | 1.069 | 0.814 |
| 578 | -0.002 | 0.042 | 0.065 | 0.471 | 0.755 | 0.869 | 0.862 | 0.79 | 1.083 | 0.858 |
| 613 | -0.001 | 0.042 | 0.106 | 0.441 | 0.778 | 0.882 | 0.871 | 0.825 | 1.018 | 0.873 |
| 689 | -0.001 | 0.045 | 0.073 | 0.44 | 0.783 | 0.885 | 0.881 | 0.815 | 1.019 | 0.857 |
| 708 | -0.001 | 0.042 | 0.118 | 0.505 | 0.72 | 0.849 | 0.84 | 0.76 | 1.042 | 0.818 |
| 744 | -0.0 | 0.042 | 0.117 | 0.504 | 0.722 | 0.85 | 0.843 | 0.757 | 1.014 | 0.808 |
| 780 | -0.001 | 0.04 | 0.075 | 0.432 | 0.794 | 0.891 | 0.889 | 0.821 | 1.035 | 0.862 |
| 803 | 0.001 | 0.043 | 0.137 | 0.512 | 0.711 | 0.845 | 0.842 | 0.733 | 0.962 | 0.786 |
| 949 | -0.001 | 0.039 | 0.064 | 0.459 | 0.769 | 0.877 | 0.875 | 0.807 | 1.037 | 0.856 |
| 999 | 0.001 | 0.045 | 0.109 | 0.429 | 0.799 | 0.895 | 0.89 | 0.812 | 0.961 | 0.849 |
| 1025 | -0.001 | 0.04 | 0.079 | 0.466 | 0.763 | 0.874 | 0.869 | 0.804 | 1.041 | 0.855 |
| 1064 | 0.001 | 0.044 | 0.11 | 0.43 | 0.797 | 0.894 | 0.889 | 0.803 | 0.955 | 0.84 |
| 1124 | 0.0 | 0.034 | 0.076 | 0.487 | 0.737 | 0.858 | 0.847 | 0.808 | 0.986 | 0.871 |
| 1146 | -0.001 | 0.032 | 0.057 | 0.46 | 0.769 | 0.877 | 0.868 | 0.826 | 1.033 | 0.881 |
| 1174 | 0.001 | 0.033 | 0.079 | 0.469 | 0.756 | 0.87 | 0.855 | 0.806 | 0.963 | 0.861 |
| 1177 | 0.001 | 0.039 | 0.068 | 0.424 | 0.801 | 0.896 | 0.89 | 0.829 | 0.978 | 0.866 |
| 1214 | 0.0 | 0.032 | 0.073 | 0.477 | 0.75 | 0.866 | 0.852 | 0.81 | 0.978 | 0.867 |
| 1217 | 0.001 | 0.035 | 0.098 | 0.482 | 0.746 | 0.864 | 0.839 | 0.799 | 0.945 | 0.863 |
| 1260 | 0.0 | 0.037 | 0.052 | 0.424 | 0.801 | 0.895 | 0.89 | 0.84 | 0.994 | 0.879 |
| 1296 | 0.0 | 0.03 | 0.063 | 0.463 | 0.766 | 0.876 | 0.862 | 0.815 | 0.98 | 0.865 |
| 1327 | 0.001 | 0.034 | 0.09 | 0.493 | 0.737 | 0.859 | 0.821 | 0.797 | 0.953 | 0.862 |
| 1442 | 0.001 | 0.032 | 0.078 | 0.478 | 0.754 | 0.869 | 0.843 | 0.778 | 0.936 | 0.833 |

## Appendix B: Statistical models detailed explanation

### B1 Linear regression

In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables, respectively). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression (multi-linear). This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data (this is a machine learning strategy, (Hastie et al., 2001)). Such models are called linear models. Most commonly, the conditional mean of the response (or predictand) given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used (see generalized linear models, GLMs at Hastie et al. (2001)). Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis.

For our case study, the dependent variable (storm surge maxima) will be reconstructed given a set of independent variables (atmospheric predictor), and the $\beta$ coefficients in the equation below represent the coefficients that are inferred from the data:

$$y = \mathbf{X}\beta + \varepsilon \text{ , where: } y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} \mathrm{x}_1^T \\ \mathrm{x}_2^T \\ \vdots \\ \mathrm{x}_n^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}. \tag{B1}$$

where $p$ relates to the number of independent variables (columns in the 2-dimensional dataset) and $n$ is the number of data points or rows in the dataset. Given this equation, the optimal coefficients can be calculated so the quadratic sum of errors is minimized:

$$\beta_{opt} = \arg\min_{\beta} \sum_{i=1}^{n} \left(\beta \cdot x_i - y_i\right)^2 \rightarrow \beta_{opt} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y \tag{B2}$$

where once again, the index $i$ refers to each data point or row in the data. Notice that to prove that the $\beta$ obtained is indeed the local minimum, one needs to differentiate once more to obtain the Hessian matrix and show that it is positive definite. This is provided by the Gauss–Markov theorem, see (Hastie et al., 2001) and (Wikipedia).

The principal benefit of this methodology is clear given its definition, as long as its coefficients are linear, the optimal coefficients are deterministic, leaving no room to convergence issues. Moreover, the time this model takes to calculate these coefficients is minimal, although it requires a sufficient RAM capacity to invert the big data matrices.

### B2 $k$-NN regression

Another machine learning tool used is the $k$-nearest-neighbors algorithm ($k$-NN), which is a non-parametric classification/regression method first developed by Evelyn Fix and Joseph Hodges in 1951, see (Fix and Hodges, 1989), and later expanded by Thomas Cover, see (Altman, 1992).

In $k$-NN regression the output is the property value for the object. This value is the average of the values of $k$-nearest neighbors (although also a weighted sum might be applied). $k$-NN is a type of regression where the function is only approximated locally and all computation is deferred until function evaluation. Since this algorithm relies on distance, if the features represent different physical units or come in vastly different scales then normalizing the training data can improve its accuracy dramatically, but in this case, the principal components are not normalized, as usually the first PCs have values much larger than the following ones, and are just these ones which we are more interested in.

In this line, the training examples are vectors in a multidimensional feature space, each with a predicted scalar value. The training phase of the algorithm consists only in storing the feature vectors and predicted values of the training samples. Then, the validation process will evaluate how the reconstruction of the test data differs from the original storm surge values, where the $k$-NN will be the nearest neighbors to the new data point, with respect to a distance metric, which in our case is the euclidean distance. This model has different hyperparameters that can be fine-tuned, as the already mentioned distance metric, which might indeed require several parameters, or the weights each closest neighbor might get to calculate the final weighted value, if this is required, but the hyperparameter with the most influence in the model performance is the number of neighbors used (Altman, 1992; Fix and Hodges, 1989; Gutiérrez et al., 2004).

The number of neighbors is crucial, and depending on the problem, the way to find the optimal value of this parameter might differ. In this study, we evaluate the models performance with the Modified Kling-Gupta Efficiency, see (Kling et al., 2012), explained in the results section, as the behavior of our models in the extreme values is very important. In this context, the less neighbors are chosen, the better the model reconstructs the extreme events in the historical dataset, but more bias is introduced in the predictions. This is known as the bias-variance trade-off, and is one of the main concepts of discussion in machine learning nowadays.

## B3 Gradient boosting regression

Another used statistical method is gradient boosting, which builds an additive model in a forward stage-wise fashion that allows for the optimization of an arbitrary differentiable loss function (usually the mean squared error, which is usually divided by two to simplify its derivative), see (Friedman, 2000). At each stage, a regression tree is fitted on the negative gradient of the given loss function, so it exploits the capabilities of several regression trees together, but using them so a loss function is optimized, and then the final output is optimal.

Gradient boosting is a machine learning technique used in regression and classification tasks. It gives a prediction model in the form of an ensemble of weak prediction models (ensemble methods), which are typically decision trees. When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms random forest and Adaboost, see (Friedman, 2000), where the ensemble is only used to calculate the final mean off all the pre-trained trees, and the weak learners are added in series, respectively.

To understand gradient boosting, decision trees must be also explained. Tree-based methods partition the feature space into a set of rectangles, and then fit a simple model (like a constant) in each one. They are conceptually simple yet powerful. Let's consider a regression problem with continuous response $y$, which is the storm surge signal, and inputs $\mathbf{X}$, which are the PCs of the atmospheric predictor. Figure B1 shows a partition of the feature space where just the first and the second PCs are taken into account, and how the final $f(\mathbf{X})$ should look like if the mean value at each final leaf is calculated. This partitions of the feature space are calculated so this previously mentioned metric is minimized, although different criterions might be used.

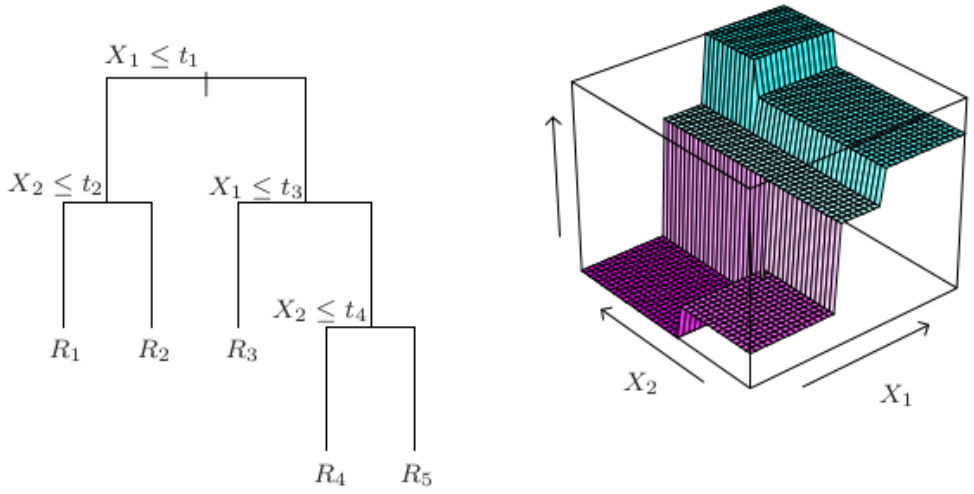


**Figure B1.** The final structure of an individual "weak" learner / decision tree is shown. In the left, the partitions depending on two input variables are shown, and in the right, the final $f(\mathbf{X})$ surface, reconstructing the target values depending on $X_1$ and $X_2$ can be also depicted.

As it is shown in the figure, the output of the model can be explained with Eq.B3:

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m) \tag{B3}$$

where the output of the model for each point $x$ will be the mean of all the target values of the group $x$ belongs to, if the mean squared error divided by 2 is minimized, which is the case.

Now, like other boosting methods, gradient boosting combines these weak "learners" into a single strong learner in an iterative way. It is easiest to explain in the least-squares regression setting, where the goal is to "teach" a model to predict values of the form $\hat{y} = f(x)$ by minimizing the mean squared error $\frac{1}{p}\sum_i (\hat{y}_i - y_i)^2$, where $i$ indexes over some training set of size $n$ of actual values of the output variable $y$, i.e., the storm surge.

Now, let us consider a gradient boosting algorithm with $M$ stages / iterations. At each stage $m$ of gradient boosting, suppose some imperfect model $F_m$ (for low $m$, this model may simply return $\hat{y}_i = \bar{y}$). In order to improve $F_m$, our algorithm should add some new estimator, $h_m(x)$. Thus,

$$F_{m+1}(x) = F_m(x) + h_m(x) = y \text{ or } h_m(x) = y - F_m(x) \tag{B4}$$

Therefore, gradient boosting will fit $h$ to the residuals. As in other boosting variants, each $F_{m+1}$ attempts to correct the errors of its predecessor $F_m$. Finally, we end up with $M$ decision trees, which as a group, outperforms the capabilities of individual "weak" trees.